FROM PARALLEL TO SEQUENTIAL: SMALL-BATCH NOISE RESTORES SIMPLICITY BIAS OF MUON

Anonymous authors
Paper under double-blind review

ABSTRACT

Simplicity bias, which can arise from sequential feature learning during training, has been proposed as a general explanation for the success of deep learning. An important driver of this phenomenon is saddle-to-saddle behavior, in which the gradient flow trajectory remains close to a sequence of saddle manifolds while learning one feature at a time. However, Muon, a state-of-the-art optimizer based on spectral descent, does not exhibit such dynamics in the full-batch setting and instead learns features in parallel. How is it then that Muon is so successful in deep learning settings? We show that stochastic gradient noise fundamentally changes this behavior: small-batch noise restores sequential feature learning, providing a mechanism for simplicity bias that is absent from full-batch Muon. We explain this transition through stochastic smoothing of the spectral descent operator underlying Muon. When gradients are small relative to the noise scale, the expected spectral update becomes approximately proportional to the actual gradient, which restores gradient-flow-like dynamics. We formalize this mechanism for deep linear models by deriving the effective continuous dynamics of noisy spectral descent. Experiments on low-rank matrix reconstruction, vision transformers, and large language models (LLMs) support these findings.

1 INTRODUCTION

Deep neural networks are heavily overparameterized: Many different predictors can fit the training data, yet, among the many existing solutions optimization often selects ones that generalize well. One proposed explanation is a simplicity bias: optimization preferentially learns simpler predictors. Sequential feature learning can produce such a bias when simpler features are learned before more complex ones. At the level of the optimization trajectory, this behavior is often described through saddle-to-saddle dynamics, where the trajectory moves through a sequence of increasingly complex regions of parameter space (Pesme & Flammarion, 2023; Boursier et al., 2022; Saxe et al., 2014).

This becomes less clear once we go beyond gradient descent. Muon Jordan et al. (2024), a state-of-the-art optimizer based on spectral descent (Bernstein & Newhouse, 2025), can behave very differently from gradient flow (Dragutinovi'c et al., 2026; Rhee et al., 2026). In simple matrix-factorization problems, gradient flow learns singular directions sequentially, whereas full-batch Muon learns them in parallel. This is surprising: if sequential feature learning is an important source of simplicity bias, why does an optimizer that removes this behavior work so well in modern deep learning?

In this work, we show that the missing piece in this consideration is small-batch noise. The behavior of Muon changes qualitatively once the gradients become stochastic. While full-batch spectral descent learns different features in parallel, small-batch spectral descent can recover a clear ordering in which stronger features are learned before weaker ones. By learning dominant singular directions first, these dynamics favor low-rank intermediate predictors, recovering a mechanism for simplicity bias that is absent from deterministic Muon.

Muon applies a matrix sign, or spectral normalization, to the gradient. In the deterministic setting, this removes much of the information contained in the gradient magnitude and causes different singular directions to evolve at the same rate. Our main insight is that small-batch noise smoothens this nonlinear sign operation. When the gradient is small relative to the noise scale, the expected

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

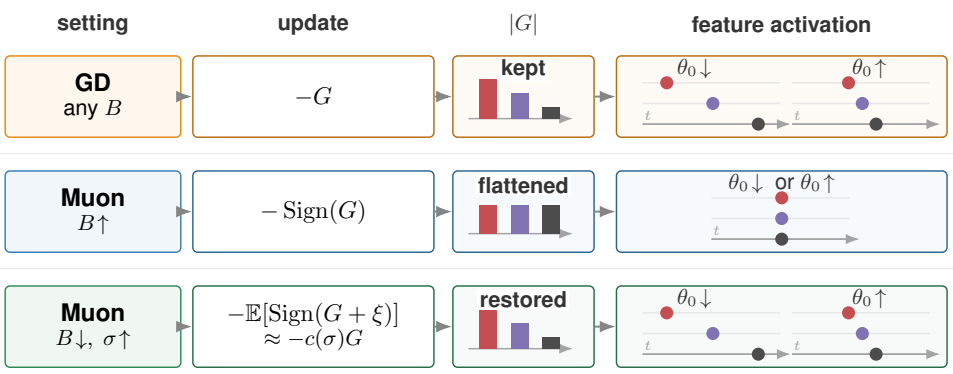


Figure 1: **Batch size B controls the simplicity bias of Muon.** Gradient descent preserves mode-dependent gradient magnitude ($|G|$), producing sequential feature activation from small initialization. In contrast, full-batch Muon ($B\uparrow$) flattens these magnitudes and learns modes in parallel, whereas small-batch noise ($\sigma\uparrow$) smooths the spectral sign operator and restores initialization-dependent sequential learning.

normalized update becomes approximately proportional to the original gradient. As a result, the gradient magnitude enters the dynamics and larger target features are learned faster. We schematically illustrate this in Figure 1.

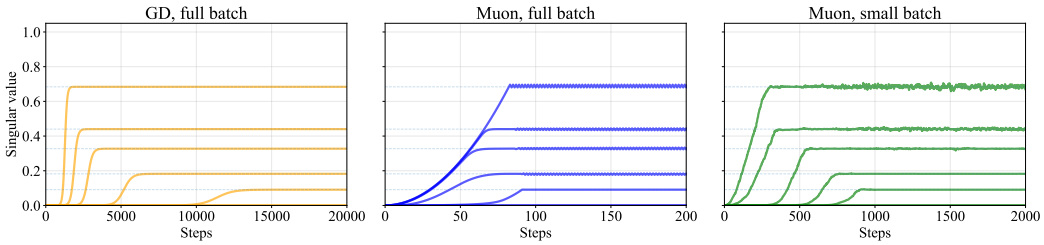


Figure 2: For low-rank matrix reconstruction with whitened data, gradient flow learns singular directions sequentially, while full-batch Muon (without momentum) learns them in parallel. Small-batch noise restores sequential feature learning, thereby giving a mechanism for simplicity bias in Muon.

We derive this mechanism explicitly for deep linear matrix factorization. For whitened data, we first obtain the singular-value dynamics of deterministic spectral flow (the flow proxy for full-batch Muon) and show that it produces parallel feature learning. We then introduce loss-dependent stochastic gradient noise and derive the corresponding expected spectral flow (i.e. the flow proxy for small-batch Muon). The resulting smoothed dynamics exhibit a gradient dependent growth rate, which leads to separation between singular modes and sequential feature learning. Increasing depth further amplifies this trend. The expected spectral flow corresponding to Muon retains this leading-order smoothing mechanism even for symmetric heavy-tailed noise and even with infinite variance, which could substantially change dynamics of SGD. For the diagonal restriction of the dynamics, we further show that depth changes the implicit bias: under infinitesimal initialization, the induced regularization is of $L_{1+1/L}$-type at depth L, whereas gradient flow retains an L_1 bias. The corresponding mirror potential also remains convex at arbitrary depth and is related to the steepest mirror flow formulation by Jacobs et al. (2026). This contrasts with gradient flow, whose corresponding potential loses convexity beyond depth two in the balanced limit.

Our theory implies that batch size can determine whether Muon exhibits the sequential feature learning that favors low-rank intermediate predictors. This is visible, for instance, in the low-rank reconstruction example in Figure 2: Full-batch Muon without momentum (also known as spectral descent) learns the relevant singular directions in parallel, whereas decreasing the batch size produces sequential dynamics. Beyond this analytically tractable setting, we investigate whether smaller batches also favor simpler learned representations and find lower effective ranks in both vision and language models.

Contributions.

- **Small-batch noise restores a mechanism for simplicity bias in Muon.** We derive effective spectral-flow dynamics in Theorem 4.2 showing how stochastic smoothing makes the matrix-sign update approximately gradient proportional, recovering sequential feature learning from the parallel dynamics of full-batch Muon (Theorem 4.1), with depth amplifying the separation between feature-learning times.
- **Heavy-tailed noise.** We show that the same smoothing mechanism extends to heavy-tailed noise, even with infinite variance as in Example 4.7, preserving the recovery of sequential learning despite noise distributions that substantially alter SGD dynamics.
- **Implicit regularization of smoothed spectral dynamics.** In Section 5, for the diagonal restriction of the aligned factorization dynamics, we characterize the implicit bias of the smoothed sign flow underlying Muon in Theorem 5.1. The resulting depth-dependent mirror potential remains convex at arbitrary depth, unlike the corresponding potential for gradient flow.
- **A link between batch size and learned complexity.** Experiments in Section 6 show that smaller batches yield representations with lower effective rank in both vision and language models, consistent with the preference for simpler representations suggested by our theory.

2 Related Work

Sequential feature learning. A recurring observation in neural-network optimization is that features are often acquired sequentially rather than in parallel. In deep linear and matrix-factorization models, gradient dynamics can exhibit mode-dependent learning times and preferential recovery of dominant spectral components (Saxe et al., 2014; Varre et al., 2023; Huh et al., 2021; Andriushchenko et al., 2022; Gidel et al., 2019; Zhao et al., 2026). Related sequential dynamics have been characterized through trajectories that move between increasingly complex saddle manifolds in diagonal linear networks (Pesme & Flammarion, 2023), and under structured assumptions for shallow ReLU networks (Boursier et al., 2022). Complementary work distinguishes several mechanisms through which feature learning can arise in graph neural networks (Mustafa et al., 2023; Mustafa & Burkholz, 2024b;a) and in linear models (Xu & Ziyin, 2024). More broadly, incremental learning has been linked to the geometry induced by mirror flows (Berthier & Pillaud-Vivien, 2026). Most closely related to our setting, recent work proposes saddle-to-saddle dynamics as an explanation for simplicity bias across architectures (Zhang et al., 2026), while (Dragutinovi'c et al., 2026; Rhee et al., 2026) demonstrate that this behavior can disappear under full-batch Muon, where singular modes are learned substantially more in parallel. Our contribution addresses the resulting gap: We show that the absence of sequential feature learning is specifically a deterministic phenomenon, and that small-batch stochasticity can restore a strong ordering of feature-learning under Muon.

Muon, steepest descent, and optimization geometry. Muon can be viewed through steepest descent in a spectral-norm geometry, with matrix-valued gradients transformed by an approximate polar or matrix-sign operation (Bernstein & Newhouse, 2025). This places it within a broader literature relating optimization geometry to implicit bias (Gunasekar et al., 2018). Sign-based and non-Euclidean steepest-descent methods have been studied in homogeneous networks, transformers, and separable classification (Balles et al., 2020; Tsilivis et al., 2025; Li et al., 2025; Fan et al., 2025), while mirror-flow analyses characterize how parameterization and geometry can favor different representations (Pesme et al., 2024; Jacobs & Montúfar, 2026; Jacobs et al., 2026; Woodworth et al., 2020; Vaškevičius et al., 2019; Tarzanagh et al., 2023; Arora et al., 2019; Liang & Montúfar, 2025; Li et al., 2022; Azulay et al., 2021; Chou et al., 2024; 2021; 2023; Sun et al., 2022; Dominé et al., 2024; Liang et al., 2026). Related work also studies smoothed sign descent from a mirror-descent perspective (Wang & Klabjan, 2024). These results primarily concern convergence, margin or implicit bias of the limiting predictor. Instead, our focus is on the transient dynamics that determine which features emerge first. Recent modifications and analyses of spectral optimization further reveal that additional modifications can be beneficial including spectral clipping and general Schatten norms (Jiang et al., 2026; Pethick, 2026; Yang et al., 2026), while random or inverted spectral transformations can remain surprisingly competitive with the standard Muon update (Shumaylov et al., 2026). This suggests that Muon's empirical success may depend less on the precise

deterministic spectral transformation itself than on how that nonlinear transformation interacts with stochastic gradients during training.

Stochasticity and nonlinear gradient normalization. Small-batch gradient noise is generally state dependent and coupled to both the data distribution and the current loss, rather than behaving as an independent additive perturbation (Wojtowytsch, 2021; Compagnoni et al., 2025). Such stochasticity can change the solution selected by optimization, including the implicit bias of diagonal and rank-one linear networks (Pesme et al., 2021; Lyu & Zhu, 2023), and injected noise can itself induce regularization in overparameterized models (Orvieto et al., 2023). Different joint limits of learning rate, noise scale, and training time can also yield qualitatively different effective dynamics (Shalova et al., 2024). Particularly relevant to normalized methods, high-dimensional analyses of signSGD quantify how the sign operation compresses gradient noise and modifies effective preconditioning (Xiao et al., 2024), while recent Muon-specific results show that spectral normalization can improve convergence under heavy-tailed gradient noise (Hubler et al., 2026).

3 PRELIMINARIES

Matrix factorization. Deep linear networks provide a tractable setting in which gradient-flow dynamics can be studied explicitly. Let $X \in \mathbb{R}^{d\times n}$ and $Y \in \mathbb{R}^{d\times n}$ denote the input and output data matrices. For a two-layer linear network with parameters $U, V \in \mathbb{R}^{d\times d}$, we consider the empirical loss $\mathcal{L}(U,V) = \frac{1}{2n}\|UVX - Y\|_{\mathrm{F}}^2$. We assume that, $XX^\top/n = I_d$ and write $C_{yx} = YX^\top/n = P\Lambda Q^\top$ for the singular value decomposition of the input–output cross-covariance.

If the factors are initialized in the corresponding singular-vector basis, the different singular modes remain decoupled under gradient flow. We further assume balanced initialization $U_0^\top U_0 = V_0 V_0^\top$, meaning that the two factors have equal magnitude along each mode and we assume that the solution is realizable. Let θ_i denote the i-th singular value of the end-to-end matrix UV, and let λ_i be the corresponding singular value of the cross-covariance matrix C_{yx}.

Theorem 3.1 (Section 1 in (Saxe et al., 2014))**.** *Under whitened data, singular-vector alignment, and balanced initialization, each singular mode evolves independently. Up to a constant rescaling of time, its gradient-flow dynamics are*

$$\frac{d\theta_i}{dt} = \theta_i\big(\lambda_i - \theta_i\big), \qquad \theta_{i,0} = \theta_{i,\mathrm{init}}, \text{ with solution } \theta_{i,t} = \frac{\lambda_i}{1 + (\lambda_i/\theta_{i,\mathrm{init}} - 1)\, e^{-\lambda_i t}},$$

when $\lambda_i > 0$ *and* $\theta_{i,\mathrm{init}} > 0$.

The same description extends to an L-layer linear network $W_L \cdots W_1$. If the layers are aligned and balanced across depth, then the end-to-end map singular value θ_i satisfies

$$\frac{d\theta_i}{dt} = L\theta_i^{2-2/L}\big(\lambda_i - \theta_i\big).$$

Thus, depth changes the effective learning rate of each mode but does not destroy the mode-wise decoupling under the stated assumptions.

Batch noise. When we train with small batches, the effective update is a stochastic gradient g instead of the full gradient $\nabla\mathcal{L}$. In our dynamical model of Muon, we want to study the effect of such a noise on the expected dynamics. Concretely, this type of noise typically is assumed to have the following properties: 1) The noise is centered and its covariance is bounded by the loss $\mathcal{L}$ for mean squared error type of losses (Wojtowytsch, 2021). 2) The gradient estimates g are independent when conditioned on the current parameters θ. 3) The noise is coupled across layers (Pesme et al., 2021). 4) The noise has low rank. We consider a diagonal noise structure that is isotropic, so not of low rank. Concretely, we have the estimate $g_{UV} = \nabla_{UV}\mathcal{L}(UV) + \xi$, with $\xi \sim N(0, \sigma^2\mathcal{L}I)$, which satisfies the first 3 assumptions.

Steepest descent. The Muon update, without momentum, is a steepest descent step with respect to the spectral norm (Bernstein & Newhouse, 2025). The update itself is the matrix sign operator as given by Definition 3.2.

Definition 3.2 (Matrix sign operator)**.** For a matrix $G \in \mathbb{R}^{d\times d}$ and its singular value decomposition (SVD) (A, B, Λ) fulfilling $G = A\Lambda B^T$, the sign operator $\text{Sign} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ is defined as: $\text{Sign}(G) := A\,\text{diag}\big(\mathbf{1}_{\{\Lambda_{ii}>0\}}\big)\,B^T$.

For a fixed learning rate $\eta > 0$, this gives the training update $U_{k+1} = U_k - \eta\,\text{Sign}(\nabla_U \mathcal{L}(U_k V_k))$, $V_{k+1} = V_k - \eta\,\text{Sign}(\nabla_V \mathcal{L}(U_k V_k))$. Our goal is to discover the resulting dynamics and to compare them to gradient flow ($\eta \to 0$). Note that under our noise assumption, stochasticity leaves gradient flow unchanged in expectation, since $\mathbb{E}[g_{UV}] = \nabla_{UV}\mathcal{L}$. Hence, any noise induced change in the learning dynamics arises specifically from Muon's nonlinear Sign transformation.

4 Spectral Dynamics

We provide two simple solutions of the spectral dynamics underlying Muon, mirroring known results for gradient flow Saxe et al. (2014). To start, we derive the simple full batch case. Next, we present our main result characterizing the expected dynamics under noisy gradient which leads to sequential feature learning. Finally, we extend our arguments also to heavy-tailed noise (Simsekli et al., 2019).

Full-batch dynamics. We first isolate the deterministic effect of the spectral flow, which removes the gradient-magnitude information responsible for sequential feature learning. This flow is given by $dU_t = -\text{Sign}(\nabla_U \mathcal{L}(U_t V_t))dt$, $dV_t = -\text{Sign}(\nabla_V \mathcal{L}(U_t V_t))dt$.

Theorem 4.1 (Spectral flow)**.** *For whitened data, the spectral gradient flow dynamics of a singular value θ of the matrix product UV is:*

$$d\theta_t = \sqrt{\theta_t}\;\textit{Sign}\;(\lambda - \theta_t)dt, \qquad \theta_0 = \theta_{\textit{init}}.$$

This has a closed-form solution:

$$\theta_t = \left(\sqrt{\lambda} + \text{Sign}\Big(\sqrt{\theta_{\text{init}}} - \sqrt{\lambda}\Big)\left(\Big|\sqrt{\theta_{\text{init}}} - \sqrt{\lambda}\Big| - \frac{t}{2}\right)_+\right)^2$$

Proof. See Appendix C. □

The full-batch result thus removes the mode-dependent rates underlying sequential learning as shown in Figure 2.

Recovering sequential feature learning with noise. We now show that stochastic smoothing restores the direct dependence of the gradient on singular values in expectation.

Theorem 4.2 (Smoothened spectral flow)**.** *Assume whitened data, singular-vector alignment, and balanced initialization. Let $\xi_\tau \sim N\left(0, \sigma^2\mathcal{L}_\tau I\right)$. Then the expected dynamics of the i-th singular value satisfy*

$$\frac{d\theta_i}{d\tau} = \sqrt{\theta_i}\,\mathbb{E}_{N(0,\sigma^2\mathcal{L}_\tau)}\left[\textit{Sign}(\lambda_i - \theta_i + \xi_\tau)\right] = \sqrt{\theta_i}\;\text{erf}\left(\frac{\lambda_i - \theta_i}{\sqrt{2\sigma^2\mathcal{L}_\tau}}\right).$$

For large σ, under the time change $t(\tau) = \sqrt{\frac{2}{\pi\sigma^2}}\int_0^\tau \frac{ds}{\sqrt{\mathcal{L}_s}}$, the leading-order dynamics are

$$\frac{d\theta_i}{dt} = \sqrt{\theta_i}\,(\lambda_i - \theta_i)\,,\ \textit{with solution}\ \theta_t = \lambda\tanh^2\left(\frac{\sqrt{\lambda}}{2}t + \text{artanh}\sqrt{\frac{\theta_{\text{init}}}{\lambda}}\right), \tag{1}$$

when $0 < \theta_{\text{init}} < \lambda$. Moreover, the loss of the flow in (1) satisfies $\mathcal{L}_\tau \le \left(\sqrt{\mathcal{L}_0} - \frac{C}{2}\tau\right)_+^2$ with $C = \min_i \sqrt{\theta_{i,0}}\sqrt{\frac{2}{\pi\sigma^2}}$. Therefore, $\mathcal{L}_\tau = 0$ for $\tau \ge 2\sqrt{\mathcal{L}_0}/C$, showing convergence in finite time.

Proof sketch. Singular-vector alignment reduces the dynamics to scalar modes, and taking the expectation of each noisy sign update gives the error-function expression. The second step follows from a first order Taylor approximation, which remains valid for all $t \ge 0$ (see Remark 4.4). Then the ODE can be solved using a substitution. Finally, the loss bound follows from a Grönwall argument. The full proof is given in Appendix C. □

According to Theorem 4.2, larger singular values are learned earlier. Still, the flow lacks the pronounced delays observed in Figure 2, possibly because our isotropic noise model omits the low-rank structure of mini-batch gradients. Increasing depth strengthens this separation (Corollary 4.3).

Corollary 4.3 (Smoothened deep spectral flow)**.** *Consider the same setting as Theorem 4.2 with the deep matrix product of L layers. The expected smoothened spectral gradient flow dynamics after the same time rescaling are:*

$$d\theta_t = \theta_t^{1-1/L}(\lambda - \theta_t)dt, \qquad \theta_0 = \theta_{init}.$$

As $L \to \infty$, Corollary 4.3 approaches the two-layer gradient-flow dynamics of Theorem 3.1. Thus, noise and depth together recover increasingly pronounced sequential feature learning, as illustrated in Figure 3. Moreover, near small initialization, the mobility $\theta^{1-1/L}$ is less suppressive than the gradient-flow mobility $\theta^{2-2/L}$, yielding faster initial growth for all $L \geq 2$ up to a constant time rescaling.

Remark 4.4. Note that our approximation of the expected flow crucially relies on the noise construction: The proportionality to $\mathcal{L}$ makes it possible to bound the erf approximation at all time. If we had considered a fixed isotropic scale, i.e. $N(0, \sigma^2 I)$, this would have only been possible with significantly high noise whose variance would have dominated the dynamics.

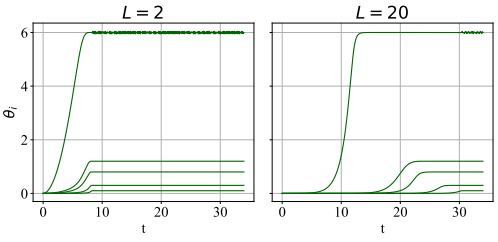


Figure 3: Simulation of the smoothened flows for depth $L = 2$ and 20. For higher depth we observe sequential feature learning.

Heavy-tailed noise. The Gaussian assumption is not needed for the expected-flow approximation. It is sufficient that the noise distribution is symmetric and sufficiently regular near the origin. This highlights another advantage of Muon over SGD, whose dynamics can change substantially under heavy-tailed gradient noise (Simsekli et al., 2019; Gurbuzbalaban et al., 2021). Since spectral descent applies a bounded sign transformation, the approximation remains valid for noise even with infinite variance. This is particularly relevant for Muon in LLM training, where heavy-tailed gradient noise has been observed. Recent work establishes convergence advantages for Muon under such noise (Hubler et al., 2026; Yu et al., 2026).

Lemma 4.5. *Let the noise ξ_i have a symmetric distribution with cumulative distribution function F_{ξ_i} and a density f_{ξ_i} that is twice continuously differentiable in a neighborhood of the origin with $f_{\xi_i}(0) > 0$. Then, for $a_i = \lambda_i - \theta_i$, the expected update as in Theorem 4.2 is:*

$$\mathbb{E}_{\xi_i}[Sign(a_i + \xi_i)] = 2F_{\xi_i}(a_i) - 1 = 2f_{\xi_i}(0)a_i + \frac{f''_{\xi_i}(0)}{3}a_i^3 + o(a_i^3).$$

Proof. By symmetry, $\mathbb{E}_{\xi_i}[\text{Sign}(a_i + \xi_i)] = 1 - 2F_{\xi_i}(-a_i) = 2F_{\xi_i}(a_i) - 1$. This follows from a Taylor expansion and the fact that symmetry implies $f'_{\xi_i}(0) = 0$. □

Remark 4.6. More explicitly, let $\xi_i = \sigma Z_i$, where Z_i has a symmetric density f_Z. Then

$$\mathbb{E}[\text{Sign}(a_i + \sigma Z_i)] = \frac{2f_Z(0)}{\sigma}a_i + \frac{f''_Z(0)}{3\sigma^3}a_i^3 + o(\sigma^{-3}).$$

Thus, as $\sigma \to \infty$, the rescaled expected dynamics recover the same gradient-proportional flow as in the Gaussian case.

Example 4.7. *Consider zero centered heavy-tailed Cauchy noise with scale parameter $\gamma > 0$. Then*

$$\mathbb{E}_{\xi_i}[Sign(a_i + \xi_i)] = \frac{2}{\pi}\arctan\left(\frac{a_i}{\gamma}\right) = \frac{2}{\pi}\frac{a_i}{\gamma} - \frac{2}{3\pi}\left(\frac{a_i}{\gamma}\right)^3 + O\left(\frac{a_i^5}{\gamma^5}\right).$$

Thus, even infinite-variance Cauchy noise preserves the smoothing mechanism underlying sequential feature learning.

Since the spectral sign update is bounded even under infinite-variance noise, stochastic fluctuations vanish as $\eta \to 0$ at fixed σ, yielding a deterministic expected flow. This contrasts with the Lévy-driven dynamics that can arise for SGD under heavy-tailed gradient noise (Nguyen et al., 2019). When $\sigma \to \infty$ jointly with $\eta \to 0$, imposing $\sigma\eta \to 0$ ensures that fluctuations also vanish on the accelerated time scale.

5 Implicit Regularization

The previous section characterized how stochastic smoothing changes the transient dynamics of spectral descent. We now characterize how the same smoothing mechanism determines the interpolator selected at convergence. For diagonal matrices, the matrix-sign operator reduces to the coordinate-wise sign. Thus, smoothened sign descent on a diagonal linear network is a direct restriction of the spectral dynamics underlying Muon. In this setting, the induced dynamics can be written as a mirror flow (see Appendix B for details). This allows us to characterize the implicit regularization that guides model selection.

Before studying the more general depth L case, for simplicity, we first consider the two-layer parameterization $\theta = u \odot v$, where $u, v \in \mathbb{R}^d$ and $\odot$ denotes pointwise multiplication. Given a dataset $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, we consider $\mathcal{L}(\theta) = \frac{1}{2n}\sum_{i=1}^n \big(\langle x_i, u \odot v\rangle - y_i\big)^2$. Unlike the whitened matrix factorization setting of Section 4, we do not require any assumption on the data covariance. Under the same noise assumptions as in the previous section, the smoothened sign dynamics of $\theta_t = u_t \odot v_t$ can (up to a positive time rescaling) be written as a mirror flow. The derivation, which relies on the preserved balance $|u_t| - |v_t| = |u_0| - |v_0|$ together with the construction of the corresponding mirror potential, is presented in Appendix D.

Theorem 5.1. *The iterates $\theta_t = u_t \odot v_t$ converge and the loss decreases with rate $\mathcal{O}(1/t^2)$ to a solution of the following objective:*

$$\min_{\theta \in \mathbb{R}^d} R(\theta, \lambda_0) \qquad \textit{such that} \qquad X\theta = y$$

where $R(\theta, \lambda_0) = \sum_{j=1}^d \left[\frac{1}{12}\left(4|\theta_j| + \lambda_0^2\right)^{3/2} - \frac{|\lambda_0|}{2}|\theta_j| - \frac{\theta_j \mathit{Sign}(\theta_{j,0})}{2}\left(\sqrt{4|\theta_{j,0}| + \lambda_0^2} - |\lambda_0|\right)\right]$ is a Legendre (Definition B.1) and Bregman (Definition B.3) potential with respect to $\theta \in \mathbb{R}^d$ and $\lambda_0 := |u_0| - |v_0|$. Moreover, if $\mathcal{L}$ satisfies with $\Lambda > 0$ the PL-inequality (Definition B.6), then for $t \geq T$ and $T = \sqrt{\frac{\pi}{2}\mathcal{L}(\theta_0)}/\left(|\lambda_0|\Lambda\right)$ convergence is guaranteed.

Proof sketch. The balance $\lambda_t := |u_t| - |v_t| = \lambda_0$ holds almost everywhere. Thus, the dynamics of θ can be written as a mirror flow with a different time clock, characterizing the optimization problem. The convergence follows from a generalized Gronwall inequality (see Appendix D).

The resulting potential is the diagonal counterpart of the square-root mobility obtained for the smoothened spectral dynamics in Section 4. Therefore, the same stochastic smoothing that restores mode-dependent learning rates also induces a non-Euclidean implicit regularization.

Remark 5.2. If $\theta_0 = 0$, then as $\lambda_0 \to 0$ we have $\lim_{\lambda_0 \to 0} R(\theta, \lambda_0) = \frac{2}{3}\|\theta\|_{L_{3/2}}^{3/2}$.

As a consequence of Remark 5.2, small balanced initialization yields an $L_{3/2}$-type implicit regularization. Thus, although stochastic smoothing restores dependence on the end-to-end loss gradient, the resulting dynamics induce a different mirror geometry from gradient flow. Restoring gradient-magnitude dependence therefore does not recover the typical L_1 bias of gradient flow (Woodworth et al., 2020). Next, we extend the result to a depth-L diagonal linear network $\theta = \Pi_{j=1}^L u_j$. As in the two-layer case, the smoothened sign dynamics preserve pairwise balances between the absolute values of the factors. These balances allow all layerwise magnitudes to be expressed in terms of θ and λ_0, producing the deep mirror geometry characterized in Jacobs et al. (2026). For the next result, we assume that the initialization satisfies the corresponding λ_0-balance across layers.

Corollary 5.3. *The iterates $\theta_t = \Pi_{j=1}^L u_{j,t}$ with $L \geq 2$, converge and the loss decreases with rate $\mathcal{O}(1/t^2)$ to a solution of the following objective:*

$$\min_{\theta \in \mathbb{R}^d} R(\theta, \lambda_0) \qquad \textit{such that} \qquad X\theta = y$$

where $R(\theta, \lambda_0)$ satisfies the following properties: 1) $R : \mathbb{R}^d \to \mathbb{R}$ is a Legendre and Bregman potential with respect to $\theta \in \mathbb{R}^d$, 2) if $\theta_0 = 0$, then there is a constant $c_L > 0$ such that

$\lim_{\lambda_0 \to 0} R(\theta, \lambda_0) = c_L \|\theta\|_{L_{1+\frac{1}{L}}}^{1+\frac{1}{L}}$. *Moreover, if* $\mathcal{L}(\theta)$ *satisfies with* $\Lambda > 0$ *the PL-inequality (Definition B.6), then for* $t \geq T$ *and* $T = \sqrt{\frac{\pi}{2}\mathcal{L}(\theta_0)} / \left(|\lambda_0|^{(L-1)}\Lambda\right)$ *convergence is guaranteed.*

Proof. See Appendix D.

Corollary 5.3 shows that depth changes the induced implicit regularization. The exponent decreases from $3/2$ at depth $L = 2$ to 1 as L increases (Table 1). Hence, increasing depth strengthens the sparsity bias of the smoothened dynamics. At the same time, the potential remains convex for every finite depth, which ensures that the corresponding implicit regularization problem remains globally well defined.

Table 1: Implicit regularization objective R for small init $\lambda_0 > 0$, depending on depth L.

	$L = 2$	$L = \infty$
SignGF	$L_{3/2}$	L_1
GF	L_1	L_1

Remark 5.4. Corollary 5.3 stands in stark contrast with the known results for gradient flow (Woodworth et al., 2020), where independent of depth L the implicit regularization is L_1 for finitely small initialization λ_0 (Table 1). Moreover, for $\lambda_0 = 0$, the corresponding gradient flow mirror potential is non-convex when $L > 2$ on $\mathbb{R}^d$, whereas the smoothened sign gradient flow mirror potential remains convex for all $L \geq 2$.

Together with Section 4, this shows that stochastic smoothing restores sequential feature learning through mode-dependent growth and induces a limiting implicit regularization approaching L_1 with increasing depth. We now examine these effects empirically.

6 Empirical validation

We validate the theory in controlled linear settings and study learned complexity in Transformers.

Matrix factorization. The matrix factorization experiment in Figure 2 illustrates our main prediction. Full-batch Muon learns singular modes in parallel, whereas small-batch noise restores sequential feature learning. In Appendix F, we provide additional ablations showing that this qualitative distinction persists when momentum is included and across a wider range of learning rates.

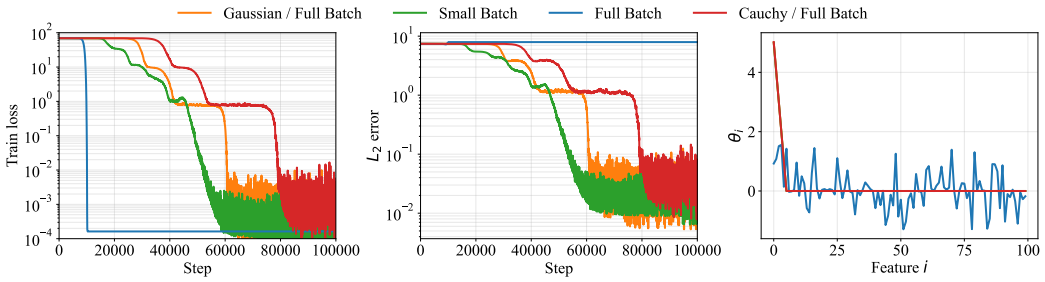


Figure 4: Noise changes the solution selected by sign descent in a diagonal linear network of depth $L = 20$. We compare full-batch updates, small-batch updates, and full-batch updates with Gaussian or Cauchy noise. Left: training loss. Middle: L_2 reconstruction error. Right: final predictor coefficients, with first 5 non-zero. Small-batch and injected-noise runs recover the same sparse target and thus largely overlap, whereas deterministic full-batch training reaches low loss without recovery.

Implicit regularization and heavy-tailed noise. We train a diagonal linear network on a sparse reconstruction task (see Appendix H). Figure 4 shows that full-batch sign descent achieves low training loss but selects a dense solution with large reconstruction error (and thus poor generalization). Small-batch updates, injected Gaussian or Cauchy noise instead recover the sparse target structure. This is consistent with the predicted $L_{1+1/L}$-type implicit regularization at large depth ($L = 20$) and indicates that the smoothing mechanism extends to heavy-tailed noise with infinite variance.

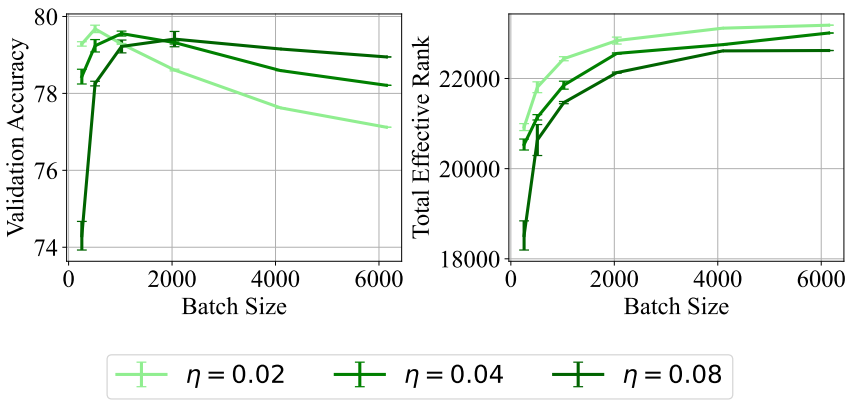


Figure 5: Final validation accuracy and effective rank of a DeiT-small model trained on ImageNet. Small batches lead to lower effective rank indicating the presence of simplicity bias.

Batch size, learning rate, and model complexity. To highlight that our general insights can transfer to transformers, we train a DeiT-small Touvron et al. (2020) on ImageNet with Muon, varying batch size and learning rate. Figure 5 shows that smaller batches yield lower effective rank (Appendix G) at the end of training across the tested learning rates, consistent with a stronger simplicity bias, while larger batches tend to reduce validation accuracy. Importantly, the same trend appears in the functionally relevant attention interaction $Q^\top K$, whose effective rank increases by roughly (42%) across the batch-size range (Figure 9). We observe similar trends on CIFAR-100 and CIFAR-10 (Appendix G). Moreover, we train a 162M-parameter Llama model on FineWeb-Edu at a fixed token budget. Table 2 shows a consistent increase in the sum of effective ranks with batch size under both fixed and linearly scaled learning rates. Together, these results support a connection between batch size and learned complexity beyond the linear setting.

Table 2: Total effective ranks for a 162M-parameter Llama model trained on 3.2B tokens from FineWeb-Edu. Larger batches ($B = 256$) yield higher effective ranks under both fixed and linearly scaled learning rates.

Batch size	Fixed η	Scaled η
B	$42{,}927.81 \pm 17.82$	$42{,}927.81 \pm 17.82$
$2B$	$43{,}040.07 \pm 15.96$	$42{,}951.67 \pm 9.63$
$4B$	$43{,}181.65 \pm 9.73$	$43{,}047.80 \pm 10.34$

7 Discussion and Conclusion

Our analysis highlights a qualitative difference between gradient flow and the spectral flow underlying Muon. Gradient flow already exhibits sequential feature learning in the deterministic setting and centered noise leaves its expected dynamics unchanged. In contrast, spectral flow removes gradient-magnitude information, causing singular modes to be learned in parallel in the full-batch setting. Stochasticity (resulting from mini-batch sampling) is therefore not just a perturbation of the deterministic Muon dynamics. By smoothing the spectral sign operator, small-batch noise reintroduces gradient-dependent growth and, together with depth, recovers sequential feature learning.

This suggests that the simplicity bias of an optimizer cannot generally be inferred from its deterministic optimization geometry alone. In our setting, stochastic smoothing affects both the transient dynamics and the selected solution, inducing a depth-dependent $L_{1+1/L}$-type implicit regularization that approaches L_1 with increasing depth. Consistent with these findings, larger batches yield higher effective rank in both vision and language models, linking batch size directly to learned complexity. More broadly, our results point to an interaction between optimizer geometry, parameterization, and stochasticity: The same deterministic update rule can exhibit qualitatively different simplicity biases depending on the training regime. Future extensions of our theory to structured mini-batch noise, nonlinear architectures, and other matrix-aware optimizers may further clarify when stochasticity merely perturbs an optimizer's implicit bias and when it fundamentally reshapes it.

AI USE STATEMENT

We used generative AI for improving the overall writing and polishing it, including the presentation of proofs. All mathematical claims and proof arguments were developed and verified by the authors. Generative AI was not used to develop theoretical models or conceptual frameworks, formulate mathematical claims, supply critical proof ingredients, propose or refine hypotheses, or design or evaluate research methodology or experiments.

REPRODUCIBILITY STATEMENT

Full experimental details and complete proofs are provided in the appendices.

REFERENCES

Felipe Alvarez, Jérôme Bolte, and Olivier Brahic. Hessian riemannian gradient flows in convex programming. *SIAM Journal on Control and Optimization*, 43(2):477–501, January 2004. ISSN 1095-7138. doi: 10.1137/s0363012902419977. URL http://dx.doi.org/10.1137/S0363012902419977.

Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Sgd with large step sizes learns sparse features. *ArXiv*, abs/2210.05337, 2022. URL https://api.semanticscholar.org/CorpusID:252815402.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/c0c783b5fc0d7d808f1d14a6e9c8280d-Paper.pdf.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 468–477. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/azulay21a.html.

Lukas Balles, Fabian Pedregosa, and Nicolas Le Roux. The geometry of sign gradient descent. *ArXiv*, abs/2002.08056, 2020. URL https://api.semanticscholar.org/CorpusID:211171365.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=hErdffTsLu.

Raphael Berthier and Loucas Pillaud-Vivien. Incremental learning in mirror flows. 2026. URL https://api.semanticscholar.org/CorpusID:289623856.

Etienne Boursier, Loucas PILLAUD-VIVIEN, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 20105–20118. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7eeb9af3eb1f48e29c05e8dd3342b286-Paper-Conference.pdf.

Vincent Chen, S.-T. Liu, Regis Cheng, Dance Yang, Shalfun Li, Ryan Yu, Lucy Liang, Hang Su, Roy Gan, Hao Wang, and Qian Wang. Dmuon: Efficient distributed muon training with near-adam overhead. *ArXiv*, abs/2606.27153, 2026. URL https://api.semanticscholar.org/CorpusID:289625072.

Hung-Hsu Chou, Johannes Maly, and Holger Rauhut. More is less: Inducing sparsity via overparameterization. *ArXiv*, abs/2112.11027, 2021. URL https://api.semanticscholar.org/CorpusID:245353663.

Hung-Hsu Chou, Holger Rauhut, and Rachel A. Ward. Robust implicit regularization via weight normalization. *ArXiv*, abs/2305.05448, 2023. URL https://api.semanticscholar.org/CorpusID:258564503.

Hung-Hsu Chou, Johannes Maly, and Dominik Stöger. How to induce regularization in linear models: A guide to reparametrizing gradient flow, 2024.

Enea Monzio Compagnoni, Rustem Islamov, Frank Norbert Proske, and Aurélien Lucchi. Unbiased and sign compression in distributed learning: Comparing noise resilience via sdes. In *International Conference on Artificial Intelligence and Statistics*, 2025. URL https://api.semanticscholar.org/CorpusID:276575195.

Francesco D'Angelo, Aditya Varre, Maksym Andriushchenko, and Nicolas Flammarion. Why do we need weight decay for overparameterized deep networks? In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023. URL https://openreview.net/forum?id=3vBtKsiHEC.

Clémentine Carla Juliette Dominé, Nicolas Anguita, Alexandra Maria Proca, Lukas Braun, Daniel Kunin, Pedro A. M. Mediano, and Andrew M Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, 2024. URL https://openreview.net/forum?id=yAE3LOjgA4.

Sara Dragutinovi'c, Yedi Zhang, and R V Ranganath. To use or not to use muon: How simplicity bias in optimizers matters. *ArXiv*, abs/2603.00742, 2026. URL https://api.semanticscholar.org/CorpusID:286222394.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. In *High-dimensional Learning Dynamics 2025*, 2025. URL https://openreview.net/forum?id=tirGweSx3a.

Gauthier Gidel, Francis R. Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in deep linear neural networks. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:140225885.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/gunasekar18a.html.

Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3964–3975. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/gurbuzbalaban21a.html.

Florian Hubler, Thomas Pethick, and Suvrit Sra. Free heavy-tailed lunch for muon: A theoretical justification of empirical success. 2026. URL https://api.semanticscholar.org/CorpusID:289263878.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. *Trans. Mach. Learn. Res.*, 2023, 2021. URL https://api.semanticscholar.org/CorpusID:232269962.

Tom Jacobs and Rebekka Burkholz. Mask in the mirror: Implicit sparsification. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=U47ymTS3ut.

Tom Jacobs and Guido Montúfar. Implicit bias of mirror flow in homogeneous neural networks: Sparse and dense feature learning. 2026. URL https://api.semanticscholar.org/CorpusID:288654339.

Tom Jacobs, Chao Zhou, and Rebekka Burkholz. Mirror, mirror of the flow: How does regularization shape implicit bias? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=MLiR9LS5PW.

Tom Jacobs, Chao Zhou, and Rebekka Burkholz. Never saddle for reparameterized steepest descent as mirror flow. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YgudIlQ9nC.

Xiaowen Jiang, Andrei Semenov, and Sebastian U. Stich. Enhancing llm training via spectral clipping. *ArXiv*, abs/2603.14315, 2026. URL https://api.semanticscholar.org/CorpusID:286571621.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:14198632.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Bingrui Li, Wei Huang, Andi Han, Zhanpeng Zhou, Taiji Suzuki, Jun Zhu, and Jianfei Chen. On the optimization and generalization of two-layer transformers with sign gradient descent. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=97rOQDPmk2.

Zhiyuan Li, Tianhao Wang, Jason D. Lee, and Sanjeev Arora. Implicit bias of gradient descent on reparametrized models: On equivalence to mirror descent. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=k4KHXS6_zOV.

Shuang Liang and Guido Montúfar. Implicit bias of mirror flow for shallow neural networks in univariate regression. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IF0Q9KY3p2.

Shuang Liang, Tom Jacobs, and Guido Montúfar. Implicit bias of sgd in multivariate relu networks: Effective width collapse. *ArXiv*, abs/2607.03613, 2026. URL https://api.semanticscholar.org/CorpusID:289934596.

Yanling Long, Yufei Gu, and Zeke Xie. Dimension-free saddle-point escape in muon. 2026. URL https://api.semanticscholar.org/CorpusID:288257409.

Bochen Lyu and Zhanxing Zhu. Implicit bias of (stochastic) gradient descent for rank-1 linear neural network. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=PjBEUTVzoe.

Nimrah Mustafa and Rebekka Burkholz. Dynamic rescaling for training gnns. *Advances in Neural Information Processing Systems 37*, 2024a. URL https://api.semanticscholar.org/CorpusID:276318335.

Nimrah Mustafa and Rebekka Burkholz. GATE: How to keep out intrusive neighbors. In *Forty-first International Conference on Machine Learning*, 2024b. URL https://openreview.net/forum?id=Sjv5RcqfuH.

Nimrah Mustafa, Aleksandar Bojchevski, and Rebekka Burkholz. Are GATs out of balance? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=qY7UqLoora.

Sree Harsha Nelaturu, Advait Gadhikar, and Rebekka Burkholz. TurboPrune: High-Speed Distributed Lottery Ticket Training. URL https://github.com/nelaturuharsha/TurboPrune.

Thanh Huy Nguyen, Umut Simsekli, Mert Gürbüzbalaban, and Gaël Richard. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:195317120.

Antonio Orvieto, Anant Raj, Hans Kersting, and Francis Bach. Explicit regularization in overparametrized models via noise injection. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7265–7287. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/orvieto23a.html.

Guilherme Penedo, Hynek Kydlícek, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. *ArXiv*, abs/2406.17557, 2024. URL https://api.semanticscholar.org/CorpusID:270711474.

Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=iuqCXg1Gng.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=vvi7KqHQiA.

Scott Pesme, Radu-Alexandru Dragomir, and Nicolas Flammarion. Implicit bias of mirror flow on separable data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=wiMaws0FWB.

Thomas Pethick. When to use what schatten-p norm in deep learning? 2026. URL https://api.semanticscholar.org/CorpusID:289301542.

Mark Rhee, James B. Simon, and Dhruva Karkada. Muon learns balanced solutions in matrix factorization without slow saddle-to-saddle dynamics. *ArXiv*, abs/2606.30509, 2026. URL https://api.semanticscholar.org/CorpusID:289684379.

Abhishek Roy, Krishnakumar Balasubramanian, Saeed Ghadimi, and Pragyya Mohapatra. Escaping saddle-point faster under interpolation-like conditions. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12414–12425, 2020.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, 2014. URL https://arxiv.org/abs/1312.6120.

Andrei Semenov, Matteo Pagliardini, and Martin Jaggi. Benchmarking Optimizers for Large Language Model Pretraining. *arXiv preprint arXiv:2509.01440*, 2025. URL https://arxiv.org/abs/2509.01440.

Anna Shalova, André Schlichting, and Mark A. Peletier. Singular-limit analysis of gradient descent with noise injection. *ArXiv*, abs/2404.12293, 2024. URL https://api.semanticscholar.org/CorpusID:269213972.

Zakhar Shumaylov, Nathael Da Costa, Peter Zaika, B'alint Mucs'anyi, Alex Massucco, Yoav Gelberg, Carola-Bibiane Schonlieb, Yarin Gal, and Philipp Hennig. Muon is not that special: Random or inverted spectra work just as well. 2026. URL https://api.semanticscholar.org/CorpusID:288258167.

Umut Simsekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. *ArXiv*, abs/1901.06053, 2019. URL https://api.semanticscholar.org/CorpusID:58028908.

Haoyuan Sun, Kwangjun Ahn, Christos Thrampoulidis, and Navid Azizan. Mirror descent maximizes generalized margin and can be implemented efficiently. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=0SVOleKNRAU.

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023. URL https://openreview.net/forum?id=gLwzzmh79K.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 2020. URL https://api.semanticscholar.org/CorpusID:229363322.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. URL https://api.semanticscholar.org/CorpusID:257219404.

Nikolaos Tsilivis, Gal Vardi, and Julia Kempe. Flavors of margin: Implicit bias of steepest descent in homogeneous neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BEpaPHDl9r.

Aditya Vardhan Varre, Maria-Luiza Vladarean, Loucas PILLAUD-VIVIEN, and Nicolas Flammarion. On the spectral bias of two-layer linear networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 64380–64414. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/cad2fd66cf88226d868f90a7cbaa4a53-Paper-Conference.pdf.

Tomas Vaškevičius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery, 2019. URL https://arxiv.org/abs/1909.05122.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Shuyang Wang and Diego Klabjan. A mirror descent perspective of smoothed sign descent. In *Conference on Uncertainty in Artificial Intelligence*, 2024. URL https://api.semanticscholar.org/CorpusID:273482612.

Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type. part i: Discrete time analysis, 2021.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/woodworth20a.html.

Ke Liang Xiao, Noah Marshall, Atish Agarwala, and Elliot Paquette. Exact risk curves of signsgd in high-dimensions: Quantifying preconditioning and noise-compression effects. *ArXiv*, abs/2411.12135, 2024. URL https://api.semanticscholar.org/CorpusID:274140980.

Yizhou Xu and Liu Ziyin. Three mechanisms of feature learning in a linear network. In *International Conference on Learning Representations*, 2024. URL https://api.semanticscholar.org/CorpusID:266999353.

Yixuan Yang, Yuqing He, and Song Li. Convergence of spectral descent for non-smooth optimization. 2026. URL https://api.semanticscholar.org/CorpusID:288672411.

Dingzhi Yu, Hongyin Tao, Yuanyu Wan, Luo Luo, and Lijun Zhang. Sign-based optimizers are effective under heavy-tailed noise. *ArXiv*, abs/2602.07425, 2026. URL https://api.semanticscholar.org/CorpusID:285453362.

Yedi Zhang, Andrew M Saxe, and Peter E. Latham. Saddle-to-saddle dynamics explains a simplicity bias across neural network architectures. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Vit5M0G5Gb.

Jiajie Zhao, Jianxing Wang, Junjie Yang, Zhiwei Bai, and Yaoyu Zhang. Gradient flow dynamics and implicit bias of diagonal linear networks under infinitesimal initialization. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=IJph1t3Egr.

A Extended related work

Saddle escape and regularization. There is also work on how stochasticity, normalization, and non-Euclidean geometry affect escape from saddle points (Jin et al., 2017; Roy et al., 2020; Long et al., 2026; Jacobs et al., 2026). In particular, dimension-independent saddle-escape guarantees have recently been established for Muon (Long et al., 2026). These results characterize local escape from saddle neighborhoods and are therefore closely connected to simplicity bias, since sequential feature learning can be viewed as progression through a hierarchy of saddles. Our work focuses on the relative escape time of each feature, which determines whether features emerge sequentially or in parallel. Finally, optimizer geometry interacts with explicit regularization and parameterization in selecting the eventual predictor (D'Angelo et al., 2023; Jacobs et al., 2025; Jacobs & Burkholz, 2025). In this work we focus however on the sequential learning and implicit regularization themselves without explicit regularization.

B Mirror flow and implicit bias

We recall the implicit regularization results for mirror flow that we use in the proofs below. Let $X \in \mathbb{R}^{n\times d}$ denote the data matrix whose i-th row is $x_i^\top$, and let $y \in \mathbb{R}^n$ denote the corresponding labels. We consider the mean squared error loss

$$\mathcal{L}(\theta) := \frac{1}{2n}\sum_{i=1}^{n}\big(\langle x_i, \theta\rangle - y_i\big)^2.$$

Given a Legendre function $R : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$, the corresponding mirror flow is defined by

$$\frac{d}{dt}\nabla R(\theta_t) = -\nabla\mathcal{L}(\theta_t), \qquad \theta_0 = \theta_{\text{init}}. \tag{2}$$

Whenever R is twice differentiable, this can equivalently be written in primal coordinates as

$$\dot\theta_t = -\nabla^2 R(\theta_t)^{-1}\nabla\mathcal{L}(\theta_t).$$

Thus, the inverse Hessian of R gives the parameter dependent mobility of the flow. This observation will allow us to identify the smoothened sign dynamics derived in the main text with a mirror flow.

Definition B.1. (Legendre Function, Definition 3.8 ((Li et al., 2022))) Let $R : \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ be a differentiable convex function. We say R is a Legendre function when the following holds: 1) R is strictly convex on the interior of its domain $\text{int}(\text{dom}R)$. 2) For any sequence $\{\theta_i\}_{i=1}^\infty$ going to the boundary of $\text{dom}R$, the gradient diverges, i.e. $\lim_{i\to\infty}||\nabla R(\theta_i)||_{L_2}^2 = \infty$.

The geometry induced by R can be expressed through its associated Bregman divergence.

Definition B.2. For a differentiable convex function $R : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$, the Bregman divergence between $\theta_1, \theta_2 \in \text{int}(\text{dom}\, R)$ is defined as

$$D_R(\theta_1, \theta_2) := R(\theta_1) - R(\theta_2) - \langle\nabla R(\theta_2), \theta_1 - \theta_2\rangle\,.$$

The Bregman divergence is generally not symmetric, but it is non-negative and vanishes when its two arguments coincide. For the Euclidean potential $R(\theta) = \frac{1}{2}\|\theta\|_2^2$, it reduces to $D_R(\theta_1, \theta_2) = \frac{1}{2}\|\theta_1 - \theta_2\|_2^2$. Moreover, adding an affine function to R changes neither its Hessian nor its Bregman divergence. To guarantee convergence of the mirror flow, we additionally use the following regularity conditions.

Definition B.3. (Bregman function Definition 4.1 (Alvarez et al., 2004)) A function R is called a Bregman function if it satisfies the following properties:

- $\text{dom}R$ is closed. R is strictly convex and continuous on $\text{dom}R$. R is C^1 on $\text{int}(\text{dom}R)$).
- For any $\theta \in \text{dom}R$ and $\gamma \in \mathbb{R}$, $\{\xi \in \text{dom}R | D_R(\theta, \xi) \le \gamma\}$ is bounded.
- For any $\theta \in \text{dom}R$ and sequence $\{\theta_i\}_{i=1}^\infty \subset \text{int}(\text{dom}R)$ such that $\lim_{i\to\infty}\theta_i = \theta$, it holds that $\lim_{i\to\infty} D_R(\theta, \theta_i) \to 0$.

Time reparameterization. The dynamics derived in the main text contain an additional positive scalar factor depending on the loss. More generally, consider

$$\dot{\theta}_t = -a_t \nabla^2 R(\theta_t)^{-1} \nabla \mathcal{L}(\theta_t), \qquad a_t > 0.$$

Under the time change $\tau(t) = \int_0^t a_s \, ds$, this becomes the standard mirror flow in Eq. (2). Consequently, such a scalar factor changes the speed at which the trajectory is traversed, but not the trajectory itself or its limiting implicit bias.

Implicit regularization. For overparameterized linear regression, the loss can admit many interpolating solutions. Mirror flow selects among them according to the Bregman geometry induced by R. The following result makes this statement precise.

Theorem B.4. *(Theorem 3.9 (Li et al., 2022)) Denote the dataset $\{x_i, y_i\}_{i=1}^n$ as (X, y). Initialized at $\theta_0 = \theta_{init}$, assume that the mirror flow Eq. (2) converges to $\theta_\infty = \lim_{t\to\infty} \theta_t$, which satisfies $X^T\theta_\infty = y$, then*

$$D_R(\theta_\infty, \theta_0) = \min_{\theta: X\theta = y} D_R(\theta, \theta_0),$$

where $D_R(\theta, \theta_0)$ is defined as in Definition B.2.

Thus, once the effective dynamics have been written as a mirror flow, their implicit regularization is determined by the corresponding potential R. The remaining condition in Theorem B.4 is convergence. We use the following result, stated as Theorem 4.14 in Li et al. (2022) and originally as Theorem 4.2 in Alvarez et al. (2004), to establish it.

Theorem B.5. *(Theorem 4.14 (Li et al., 2022)) Let R be a Legendre and Bregman function, which is twice differentiable on its interior domain and $\nabla^2 R$ is locally Lipschitz. Moreover, assume that $\mathcal{L}$ is quasi-convex, $\nabla\mathcal{L}$ is locally Lipschitz and argmin$\{\mathcal{L}(x)|x \in \mathbb{R}^n\}$ is non-empty. Then as $t \to \infty$, θ_t converges to some critical point θ^*. Moreover, if $\mathcal{L}$ is convex θ_t converges to a minimizer of $\mathcal{L}$.*

Our argument regarding the implicit regularization therefore consists of three steps: we first identify the mobility of the smoothened sign dynamics and construct the corresponding mirror potential R; we then verify that R is a Legendre and Bregman function so that Theorem B.5 guarantees convergence and finally, Theorem B.4 characterizes the selected interpolating solution. Furthermore under the additional assumption of the PL-inequality (Definition B.6) we can get convergence in finite time for our setting. Note for the MSE loss $\mathcal{L}(\theta^*) = 0$.

Definition B.6. (PL-inequality) For a differentiable function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ and positive constant $\Lambda > 0$ we say $\mathcal{L}$ satisfies the PL-inequality if for all $\theta \in \mathbb{R}^d$ we have:

$$\|\nabla\mathcal{L}(\theta)\|_2^2 \geq 2\Lambda(\mathcal{L}(\theta) - \mathcal{L}(\theta^*))$$

where $\theta^* \in \arg\min_{\theta\in\mathbb{R}^d} \mathcal{L}(\theta)$.

C PROOF OF THE SPECTRAL FLOWS

Proof of Theorem 4.1.

Proof. We first show the result for $L = 2$. The data are assumed to be whitened, $XX^\top/n = I_d$, the loss can be written, up to an additive constant independent of U and V, as

$$\mathcal{L}(U, V) = \frac{1}{2} \|UV - C_{yx}\|_{\mathrm{F}}^2, \qquad C_{yx} = P\Lambda Q^\top,$$

where

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d).$$

It follows that

$$\nabla_U \mathcal{L}(U, V) = (UV - C_{yx})V^\top$$

and

$$\nabla_V \mathcal{L}(U, V) = U^\top(UV - C_{yx}).$$

By singular-vector alignment and balanced initialization, we may write

$$U_0 = PD_0, \qquad V_0 = D_0 Q^\top,$$

where

$$D_0 = \mathrm{diag}\left(\sqrt{\theta_{1,\mathrm{init}}}, \ldots, \sqrt{\theta_{d,\mathrm{init}}}\right).$$

We first show that this structure is preserved by the spectral gradient flow. Suppose that, at some time t,

$$U_t = PD_t, \qquad V_t = D_t Q^\top,$$

where

$$D_t = \mathrm{diag}(d_1(t), \ldots, d_d(t)), \qquad d_i(t) \geq 0.$$

Then

$$U_t V_t = PD_t^2 Q^\top,$$

and hence

$$\theta_i(t) = d_i(t)^2.$$

Moreover,

$$U_t V_t - C_{yx} = P(D_t^2 - \Lambda)Q^\top.$$

Therefore,

$$\nabla_U \mathcal{L}(U_t, V_t) = P(D_t^2 - \Lambda)D_t$$

and

$$\nabla_V \mathcal{L}(U_t, V_t) = D_t(D_t^2 - \Lambda)Q^\top.$$

Define

$$S_t = \mathrm{diag}\left(\mathrm{Sign}(\theta_i(t) - \lambda_i)\right)_{i=1}^d.$$

Since all matrices inside the singular-vector bases are diagonal, the matrix sign operator satisfies

$$\mathrm{Sign}\left(\nabla_U \mathcal{L}(U_t, V_t)\right) = PS_t$$

and

$$\mathrm{Sign}\left(\nabla_V \mathcal{L}(U_t, V_t)\right) = S_t Q^\top.$$

Consequently, the spectral gradient flow is

$$\dot{U}_t = -\mathrm{Sign}\left(\nabla_U \mathcal{L}(U_t, V_t)\right) = -PS_t$$

and

$$\dot{V}_t = -\mathrm{Sign}\left(\nabla_V \mathcal{L}(U_t, V_t)\right) = -S_t Q^\top.$$

Thus, the aligned and balanced structure is preserved, and

$$\dot{D}_t = -S_t.$$

In particular, the singular modes evolve independently and satisfy

$$\dot{d}_i(t) = \text{Sign}(\lambda_i - \theta_i(t)).$$

Since $\theta_i(t) = d_i(t)^2$, we obtain

$$\dot{\theta}_i(t) = 2d_i(t)\dot{d}_i(t) = 2\sqrt{\theta_i(t)}\text{Sign}(\lambda_i - \theta_i(t)).$$

Absorbing the constant factor 2 into the time parametrization gives

$$\dot{\theta}_i(t) = \sqrt{\theta_i(t)}\text{Sign}(\lambda_i - \theta_i(t)).$$

Fixing one singular mode and suppressing the index i, we have

$$\dot{\theta}_t = \sqrt{\theta_t}\,\text{Sign}(\lambda - \theta_t), \qquad \theta_0 = \theta_{\text{init}}.$$

To solve this equation, define

$$r_t := \sqrt{\theta_t}.$$

Since $\theta_t = r_t^2$, the chain rule gives

$$\dot{\theta}_t = 2r_t\dot{r}_t.$$

Whenever $r_t > 0$,

$$2r_t\dot{r}_t = r_t\text{Sign}(\lambda - r_t^2),$$

and therefore

$$\dot{r}_t = \frac{1}{2}\text{Sign}(\lambda - r_t^2).$$

Because $r_t \geq 0$ and $\lambda \geq 0$,

$$\text{Sign}(\lambda - r_t^2) = \text{Sign}(\sqrt{\lambda} - r_t).$$

Thus, r_t moves toward $\sqrt{\lambda}$ at constant speed $1/2$ and remains there once it reaches $\sqrt{\lambda}$.

Since

$$r_0 = \sqrt{\theta_{\text{init}}},$$

the solution is

$$r_t = \sqrt{\lambda} + \text{Sign}\left(\sqrt{\theta_{\text{init}}} - \sqrt{\lambda}\right)\left(\left|\sqrt{\theta_{\text{init}}} - \sqrt{\lambda}\right| - \frac{t}{2}\right)_+.$$

Finally, using $\theta_t = r_t^2$, we obtain

$$\theta_t = \left(\sqrt{\lambda} + \text{Sign}\left(\sqrt{\theta_{\text{init}}} - \sqrt{\lambda}\right)\left(\left|\sqrt{\theta_{\text{init}}} - \sqrt{\lambda}\right| - \frac{t}{2}\right)_+\right)^2.$$

The dynamic description extension for higher depth L follows analogous, this concludes the proof. □

Proof of Theorem 4.2.

Proof. By alignment, the singular modes decouple, and Gaussian smoothing gives

$$\mathbb{E}\left[\text{Sign}(\lambda_i - \theta_i + \xi_i)\right] = \text{erf}\left(\frac{\lambda_i - \theta_i}{\sqrt{2\sigma^2\mathcal{L}_\tau}}\right).$$

Using

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}}x + O(x^3),$$

we obtain

$$\frac{d\theta_i}{d\tau} = \sqrt{\frac{2}{\pi\sigma^2}}\frac{\sqrt{\theta_i}(\lambda_i - \theta_i)}{\sqrt{\mathcal{L}_\tau}} + O\left(\frac{\sqrt{\theta_i}|\lambda_i - \theta_i|^3}{\sigma^3\mathcal{L}_\tau^{3/2}}\right).$$

Using the fact that, for all $i \in [d]$ and $\mathcal{L} > 0$, we have

$$\frac{|\lambda_i - \theta_i|}{\sqrt{2\mathcal{L}}} \leq 1,$$

we obtain a uniform bound in terms of the noise strength σ. Therefore, for sufficiently large noise strength σ, the approximation holds uniformly along the trajectory. Moreover, as the learning rate $\eta \to 0$, the stochastic fluctuations vanish, while the expected drift remains. Hence, the stated time change yields, to leading order,

$$\frac{d\theta_i}{dt} = \sqrt{\theta_i}(\lambda_i - \theta_i).$$

For a fixed singular mode, set $r_t := \sqrt{\theta_t}$. Then

$$2r_t \frac{dr_t}{dt} = r_t(\lambda - r_t^2),$$

and hence

$$\frac{dr_t}{dt} = \frac{1}{2}(\lambda - r_t^2).$$

Assuming $0 < \theta_{\text{init}} < \lambda$, separation of variables gives

$$\frac{dr_t}{\lambda - r_t^2} = \frac{1}{2}\, dt.$$

Integrating from 0 to t, and using $r_0 = \sqrt{\theta_{\text{init}}}$, yields

$$\int_{r_0}^{r_t} \frac{dr}{\lambda - r^2} = \frac{t}{2}.$$

Since

$$\int \frac{dr}{\lambda - r^2} = \frac{1}{\sqrt{\lambda}} \operatorname{artanh}\left(\frac{r}{\sqrt{\lambda}}\right),$$

we obtain

$$\operatorname{artanh}\left(\frac{r_t}{\sqrt{\lambda}}\right) = \frac{\sqrt{\lambda}}{2} t + \operatorname{artanh}\left(\sqrt{\frac{\theta_{\text{init}}}{\lambda}}\right).$$

Applying $\tanh$ to both sides gives

$$r_t = \sqrt{\lambda}\, \tanh\left(\frac{\sqrt{\lambda}}{2} t + \operatorname{artanh}\sqrt{\frac{\theta_{\text{init}}}{\lambda}}\right).$$

Finally, since $\theta_t = r_t^2$,

$$\theta_t = \lambda \tanh^2\left(\frac{\sqrt{\lambda}}{2} t + \operatorname{artanh}\sqrt{\frac{\theta_{\text{init}}}{\lambda}}\right).$$

Now we can use the time change and the fact that the solution is non-decreasing from initialization. We use the alignment and that we then can express the Frobenius norm in terms of a Euclidean norm on the singular values $\mathcal{L}_\tau = \frac{1}{2}\sum_{i=1}^{d}(\lambda_i - \theta_i)^2$, which gives us:

$$d\mathcal{L}_\tau = -\nabla\mathcal{L}_\tau \sqrt{\theta_\tau} \nabla\mathcal{L}_\tau / \sqrt{\mathcal{L}_\tau} \sqrt{\frac{2}{\pi\sigma^2}} d\tau \leq -\min_i \sqrt{\theta_{i,0}} \sqrt{\mathcal{L}_\tau} \sqrt{\frac{2}{\pi\sigma^2}} d\tau$$

Applying a Grönwall type argument gives the result.

□

D IMPLICIT BIAS

We now focus on a diagonal linear network, with the noise considered in the main text $N(0, \sigma\mathcal{L}I)$. This gives the leading order ODE:

$$\begin{cases} du_t = -\text{Sign}(v_t)\sqrt{\frac{2}{\pi}}\nabla_\theta \mathcal{L}(\theta_t)/\sqrt{\mathcal{L}t}dt \\ dv_t = -\text{Sign}(u_t)\sqrt{\frac{2}{\pi}}\nabla_\theta \mathcal{L}(\theta_t)/\sqrt{\mathcal{L}t}dt \end{cases} \tag{3}$$

Note that we now have $\mathbb{E}[\text{Sign}(v \odot g_\theta)] = \text{sign}(v) \odot \mathbb{E}[\text{Sign}(g_\theta)]$ where g_θ is the gradient estimate. In the matrix factorization the sign on singular values does not appear as we assumed there that we start at a balanced initialization. In Lemma D.1 we derive a dynamical invariant of the flow in Equation 3 for constructing of the mirror flow dynamics.

Lemma D.1. *The flow in Eq. (3) satisfies the following invariant:*

$$|u_t| - |v_t| = |u_0| - |v_0| =: \lambda_0$$

Proof. Apply the chain rule and use the flow definition in Eq. (3).

Note that Lemma D.1 also has been established for full batch case Jacobs et al. (2026).

Using Lemma D.1 we show now that the dynamics of θ follow a mirror flow by deriving the corresponding Riemannian gradient flow in Lemma D.2.

Lemma D.2. *The dynamics of $\theta_t = u_t \odot v_t$ are described by a time rescaled mirror flow:*

$$d\theta_t = -\sqrt{4|\theta_t| + \lambda_0^2}\sqrt{\frac{2}{\pi}}\nabla_\theta \mathcal{L}(\theta_t)/\sqrt{\mathcal{L}t}dt$$

Proof. It follows from the chain rule that:

$$d\theta_t = -(|u_t| + |v_t|)\sqrt{\frac{2}{\pi}}\nabla_\theta \mathcal{L}(\theta_t)/\sqrt{\mathcal{L}t}dt$$

By definition and Lemma D.1 we can use the following identities:

$$\begin{cases} \theta = u \odot v \\ \lambda_0 = |u| - |v| \end{cases}$$

This gives:

$$\sqrt{4|\theta| + \lambda_0^2} = \sqrt{4|u||v| + (|u| - |v|)^2} = \sqrt{(|u| + |v|)^2} = |u| + |v|.$$

This concludes the proof. □

Proof of Theorem 5.1. Define the a time change $\tau := \sqrt{\frac{2}{\pi}}\int_0^t 1/\sqrt{\mathcal{L}_t}dt$. Then we get the ODE with a Riemannian metric corresponding to a Legendre and Bregman potential R:

$$d\theta_\tau = -\sqrt{4|\theta_\tau| + \lambda_0^2}\nabla_\theta \mathcal{L}(\theta_\tau)d\tau.$$

It follows from Theorem 4.2 in (Alvarez et al., 2004) the iterates converge to a minimizer of L. Moreover, the loss L decreases linearly with respect to the time change τ:

$$\mathcal{L}(\theta_\tau) \leq \frac{D_R(\theta^*, \theta_0)}{\tau}$$

$$d\tau(t)/dt = \sqrt{\frac{2}{\pi}}1/\sqrt{\mathcal{L}_t} \geq \sqrt{\frac{2}{\pi}}\frac{\sqrt{\tau}}{\sqrt{D_R}}$$

which implies that

$$\tau \geq \frac{1}{2\pi}\frac{t^2}{D_R}$$

Thus

$$\mathcal{L}(\theta_t) \leq \frac{B}{t^2}$$

where $B := 2\pi D_R^2$.

Moreover, under the PL-inequality (Definition B.6) we have

$$\begin{aligned} d\mathcal{L}(\theta_t) &= -\nabla_\theta \mathcal{L}(\theta_t)\sqrt{4|\theta_t| + \lambda_0^2}\sqrt{\frac{2}{\pi}}\nabla_\theta \mathcal{L}(\theta_t)/\sqrt{\mathcal{L}_t}dt \\ &\leq -|\lambda_0|\sqrt{\frac{2}{\pi}}||\nabla_\theta \mathcal{L}(\theta_t)||_2^2/\sqrt{\mathcal{L}(\theta_t)}dt \\ &\leq -2|\lambda_0|\sqrt{\frac{2}{\pi}}\Lambda\sqrt{\mathcal{L}(\theta_t)}dt \end{aligned}$$

Therefore by a generalized Grownwall bound we have:

$$\mathcal{L}(\theta_t) \leq \left(\sqrt{\mathcal{L}(\theta_0)} - |\lambda_0|\sqrt{\frac{2}{\pi}}\Lambda t\right)_+^2.$$

Finally, the KKT condition follows from the mirror flow construction. Concretely, the flow for θ_t is equivalent to a mirror flow with $\nabla_\theta R(\theta)$ such that:

$$d\nabla_\theta R(\theta_t) = -\nabla_\theta L(\theta_t)dt, \qquad \theta_0 = \theta_{\text{init}},$$

by definition of the loss and integrating over time we have that for all $t \geq 0$

$$\nabla_\theta R(\theta_t) = -X^T \int_0^t (X\theta_s - y)ds \in \text{span}(X),$$

which is exactly the KKT condition of the optimization problem. □

Proof of Corollary 5.3. The argument follows the proof of Theorem 5.1. The smoothened sign dynamics preserve the pairwise balances $|u_{j,t}| - |u_{1,t}| = \lambda_0$ almost everywhere. Consequently, after the same time change as in Theorem 5.1, the product $\theta_t = \Pi_{j=1}^L u_{j,t}$ satisfies

$$\frac{d\theta_t}{dt} = -M_{L,\lambda_0}(\theta_t)\nabla\mathcal{L}(\theta_t),$$

where M_{L,λ_0} is the separable inverse metric induced by these balances. This is precisely the $q = 1$ metric characterized in Jacobs et al. (2026, Theorem 4.8). Since $(L-1)/L < 1$, the corresponding potential $R(\theta, \lambda_0)$ is a Legendre and Bregman function for every $L \geq 2$. The mirror-flow implicit-bias and convergence results from Appendix B therefore apply. Repeating the argument of Theorem 5.1 gives $\mathcal{L}(\theta_t) = \mathcal{O}(t^{-2})$ and shows that the limiting interpolator minimizes $R(\theta, \lambda_0)$ subject to $X\theta = y$.

For $\lambda_0 \to 0$, the explicit characterization in Jacobs et al. (2026, Lemma 4.13), specialized to $q = 1$, gives, up to irrelevant affine and positive multiplicative terms,

$$R(\theta, 0) \propto \sum_{i=1}^d |\theta_i|^{1+1/L} = \|\theta\|_{1+1/L}^{1+1/L}.$$

Finally, Jacobs et al. (2026, Corollary 4.9) gives the coercivity bound $M_{L,\lambda_0}(\theta) \succeq \lambda_0^{L-1} I$. Under the PL inequality (Definition B.6), the same differential inequality as in Theorem 5.1 therefore yields

$$T = \sqrt{\frac{\pi}{2}}\frac{\sqrt{\mathcal{L}(\theta_0)}}{\lambda_0^{L-1}\Lambda}.$$

□

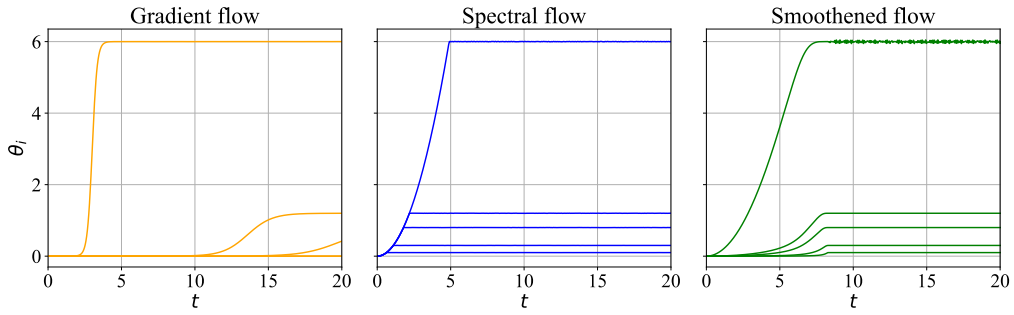


Figure 6: Simulated flows for all 3 settings at depth $L = 2$: gradient, spectral, smoothened smoothened flow.

E Flow simulation details

To illustrate that our derived flow description captures the underlying dynamics we simulate the flows derived in the main text. We set the non-zero singular values to $\lambda := [6.0, 1.2, 0.8, 0.3, 0.1]$.

We set $L = 2$ in the first experiment and later compare $L = 2, 20$ for the smoothened flow to illustrate the sequential feature learning emerging from depth. For $L = 2$ we can make use of the explicit solutions derived and we use them for the smoothened and smoothened delayed flows. The initialization scale is $\theta_{\text{init}} = 1e-4$. For the other flows, we use Runge-Kutta 45 numerical integrator from the SciPy library to simulate the flow (Virtanen et al., 2020).

We observe in Figure 6 that gradient flow indeed exhibits sequential feature learning and that the spectral flow learns all features in parallel. For the smoothened flow we observe that there is a growth separation. However, all features are learned at the same time. In the main text we also consider the smoothened flow and compare depths in Figure 3, where do observe sequential feature learning emerge for higher depth.

F Matrix factorization details

Our theoretical analysis studies a simplified matrix factorization problem. To show that our flows derived capture the dynamics in this matrix factorization problem we run the actual matrix factorization problem with small batch gradients and full batch gradients. We run gradient descent and Muon with learning rate $\eta = 1e-2$. For all this toy experiment we set momentum $\beta = 0$. We run the experiments for 50000 steps. We use whitened data, we set the dimension to of the matrix to $d = 10$ and use $n = 2000$ data points. The rank of the teacher model is set to 5. We consider for gradient descent and Muon full batch and for Muon in addition the small batch setting with one sample. Note we still run of 50000 steps not epochs. We use a small initialization, with $U \sim \alpha N(0, I)$ and $V \sim \alpha N(0, I)$ where $\alpha = 1e-7$ for three-way comparison and for the depth comparison it $\alpha = 1e-4$. Finally, we track in Figure 2 in the main text the evolution of the 5 non-zero singular values.

Furthermore, we conduct an additional experiment to include momentum and keeping the other hyperparameters the same. The results are presented in Figure 7. Moreover, we conduct an experiment for varying the learning rate and show that this does not significantly alter the flow as demonstrated in Figure 8.

G Vision and language experimental details

To understand the effect of batch size we run standard training recipes for Muon on a small DeiT for CIFAR-10. The codebase implementation is based on Nelaturu et al.. We conducted an exhaustive grid search over the learning rate and batch size and each configurations is ran for 3 seeds for which we report the mean and standard 95% confidence intervals. The batch size values in the sweep refer



Figure 7: Gradient descent, Muon with small and large batches. The sequential learning phenomenon persists with momentum included.

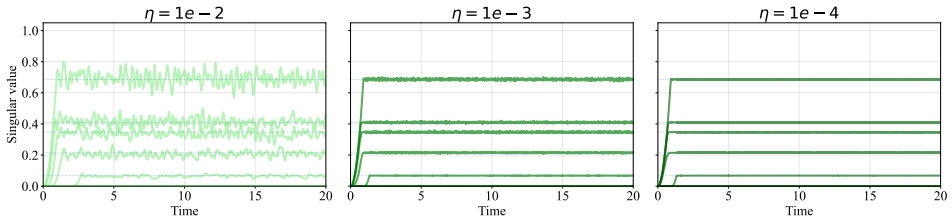


Figure 8: Matrix factorization varying the learning rate for small batch Muon. We observe that the trajectories remain similar to each other accounting for discretization time.

to the batch size used on each individual GPU and were selected from

$$\{32, 64, 128, 256, 512, 1024, 2048, 4096\}.$$

All experiments were conducted using distributed training across 4 GPUs, resulting in effective global batch sizes ranging from 128 to 16384 samples. The learning rate was independently varied over

$$\eta \in \{0.005, 0.01, 0.02\},$$

yielding a total of 24 training configurations. Each configuration was trained for 300 epochs using the same model architecture, optimizer (Muon), scheduler (TriangularSchedule with 10 warmup steps), weight decay set to 0.01, CutMix, and MixUp. For CIFAR 100, we used the exact same configuration sweep and hyperparameters.

For ImageNet we train a DeiT-small and use Label smoothening for the loss. The batch size was set to

$$B \in \{64, 128, 256, 512, 1024, 1536\} * 4.$$

The learning rate was set to

$$\eta \in \{0.02, 0.04, 0.08\}.$$

Moreover, we only run one seed for the two largest batch configurations i.e. per GPU: 1024 and 1536.

The results for CIFAR-10 and CIFAR-100 are reported in Figures 10 and 11. We observe that simplicity bias correlates with better performance overall.

Language experiment. We conduct the same experiment in the language domain. We consider the pretraining a LLama-type model Touvron et al. (2023) on the FineWeb-Edu data set Penedo et al. (2024) for $3.2B$ tokens. For this experiment we use the codebase Semenov et al. (2025). As in the vision experiment we vary the batch size and learning rate and run for 3 seeds for which we report the mean and standard 95% confidence intervals.

The model we consider has 12 layers, embedding dimension 768, 12 heads, and untied embeddings. We train using D-Muon Chen et al. (2026) with a WSD learning-rate schedule and sequence length 512. We vary the batchsize but we also use a fixed gradient accumulation of 32. This amounts to $32/4 = 8$ gradient accumulations for each A100 GPU. For training, we use a base learning rate of $\eta = 1e-3$, momentum parameters $\beta_1 = 0.8$ and $\beta_2 = 0.999$ for the additional modules trained with Adam Kingma & Ba (2017), gradient clipping at 0.5.

Table 3: Validation accuracy for a 162M-parameter Llama model trained on 3.2B tokens from FineWeb-Edu ($B = 256$). Results report means and approximate 95% confidence intervals over three seeds.

Batch size	Fixed η	Scaled η
256	40.79 ± 0.07	40.79 ± 0.07
512	40.48 ± 0.06	40.71 ± 0.09
1024	40.18 ± 0.02	40.78 ± 0.05

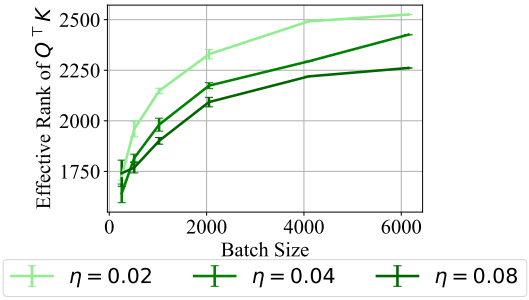


Figure 9: Effective rank of QK for the DeiT-small model trained on ImageNet.

For the batch size, including the gradient accumulation, we use $B \in \{256, 512, 1024\}$. We adapt the amount of iterations to the batch size to match the amount of tokens seen. For $B = 1024$ we use 6104.

In Table 3 we report the corresponding validation accuracy of the reported effective ranks in the main text. We observe that there is a slight performance degradation when keeping the learning rate fixed and increasing the batchsize while rescaling it keeps the performance similar.

Effective rank and norm ratio. We furthermore tracked the effective rank of each linear layer by computing it singular values σ, and nummerical stability constant $\epsilon = 1e-12$:

$$p = \frac{\sigma}{\sum_i \sigma_i}, \qquad \text{EffRank}(p) := \exp(-\sum_i p_i \log(p_i + \epsilon)).$$

In the main text we report the total sum of the effective ranks of each linear layer of the model. For the Imagenet experiment we also report the effective rank of $Q^\top K$ in Figure 9, where we observe the same pattern.

H Diagonal linear network

In this section we provide details for the diagonal linear network experiment. We consider two depth configurations $L = 2, 20$. Moreover we compare sign descent with gradient descent. We also incorporate noise injections both Normal and Cauchy. The noise is added directly to the gradient with respect to θ i.e. the shared gradient. We sample random normally distributed data with dimension $d = 100$ and $n = 40$ samples. The ground truth is set to a sparse vector $\theta = (5, 4, 3, 2, 1, 0, \ldots, 0)$.

For sign descent and depth $L = 20$, we use initialization scale $\lambda_0 = 1.11e-2$, learning rate $\eta = 1e-4$ and train for iterations $T = 100000$. For small batch, we set the batch to 1 and the noise injection is done by sampling i.i.d. from the respective noise distributions and scaling with $\sigma = 3\sqrt{f}$, where f is the mean squared error loss. The result is reported in the main text in Figure 4

For sign descent and depth $L = 2$, we use initialization scale $\lambda_0 = 1.11e-3$, learning rate $\eta = 1e-5$ and train for iterations $T = 2000000$. For small batch, we set the batch to 1 and the noise injection is done by sampling i.i.d. from the respective noise distributions and scaling with $\sigma = 3\sqrt{f}$, where f is the mean squared error loss. The result is reported in Figure 12. We observe that the noisy configurations reach a simpler solution than full batch but do not recover the exact solution. This is in line with an $L_{3/2}$ bias.

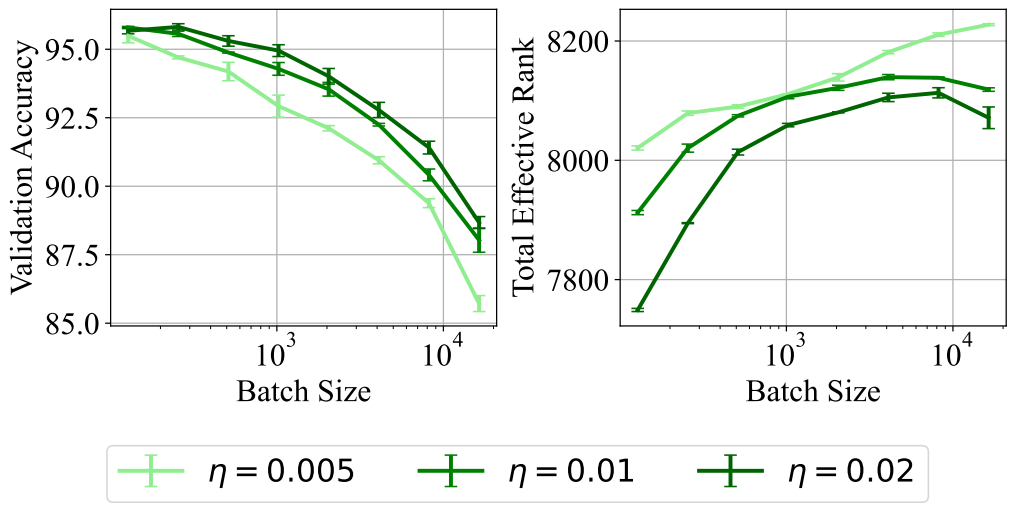


Figure 10: Final validation accuracy and effective rank of a small DeiT model trained on CIFAR 10. Small batches lead to lower rank indicating the presence of simplicity bias.

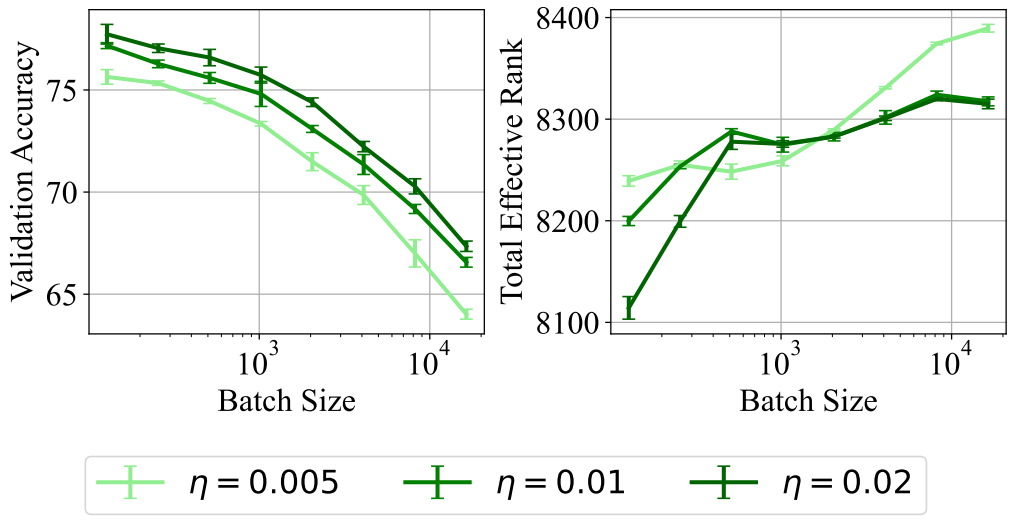


Figure 11: Final validation accuracy and summed norm ratio of a small DeiT model trained on CIFAR-100. Small batches leads to lower rank indicating the presence of simplicity bias.

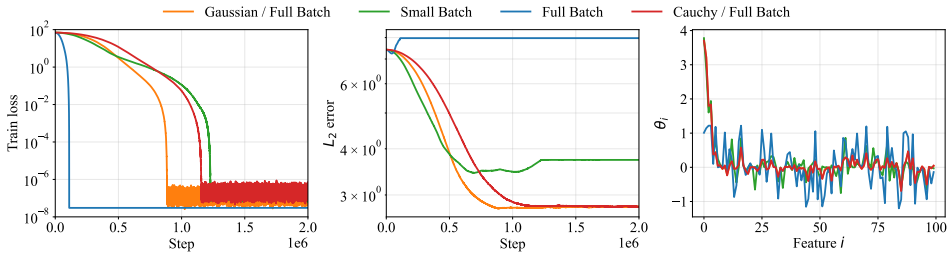


Figure 12: Noise changes the solution selected by sign descent in a diagonal linear network of depth $L = 2$. We compare full-batch updates, small-batch updates, and full-batch updates with Gaussian or Cauchy noise. Left: training loss. Middle: L_2 reconstruction error. Right: final predictor coefficients, with first 5 non-zero. Small-batch training and both forms of injected noise recover the sparse target better but not full which is in line with the $L_{3/2}-$bias, whereas deterministic full-batch training achieves low training loss without recovery.

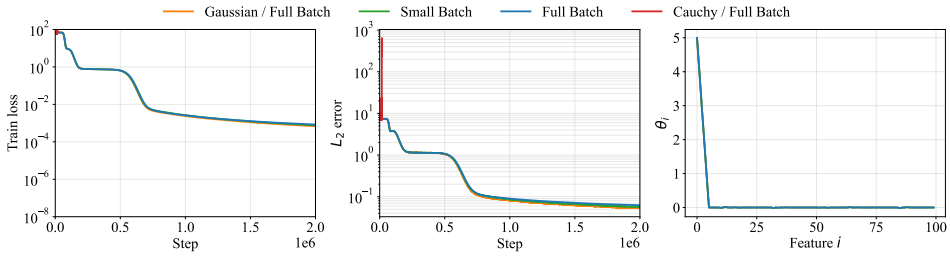


Figure 13: Noise does not change the solution selected by gradient descent in a diagonal linear network of depth $L = 2$ for small initialization. We compare full-batch updates, small-batch updates, and full-batch updates with Gaussian or Cauchy noise. Left: training loss. Middle: L_2 reconstruction error. Right: final predictor coefficients, with first 5 non-zero. All recover the ground truth solution except the injected heavy tail noise due to too large variance.

For gradient descent and depth $L = 2$, we use initialization scale $\lambda_0 = 1.11e - 3$, learning rate $\eta = 1e - 5$ and train for iterations $T = 2000000$. For small batch, we set the batch to 1 and the noise injection is done by sampling i.i.d. from the respective noise distributions and scaling with $\sigma = 3\sqrt{f}$, where f is the mean squared error loss. The result is reported in Figure 13. Observe that the heavy tail noise injection made the dynamics unstable. The other configurations exhibit the expected gradient flow saddle-to-saddle dynamics Pesme & Flammarion (2023).

Note that training gradient descent for high depth takes an extraordinary amount of time due to the mobility being super linear.